Received: 20 July 2020 | Revised: 21 September 2020 | Accepted: 23 September 2020

DOI: 10.1002/jcc.26432

FULL PAPER

Journal of COMPUTATIONAL CHEMISTRY WILEY

ProteinUnet—An efficient alternative to SPIDER3-single for sequence-based prediction of protein secondary structures

Krzysztof Kotowski[1] | Tomasz Smolarczyk[1] | Irena Roterman-Konieczna[2] | Katarzyna Stapor[1]

[1]Department of Applied Informatics, Silesian University of Technology, Gliwice, Poland

[2]Department of Bioinformatics and Telemedicine, Jagiellonian University Medical College, Kraków, Poland

Correspondence
Katarzyna Stapor, Department of Applied Informatics, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland.
Email: katarzyna.stapor@polsl.pl

Abstract

Predicting protein function and structure from sequence remains an unsolved problem in bioinformatics. The best performing methods rely heavily on evolutionary information from multiple sequence alignments, which means their accuracy deteriorates for sequences with a few homologs, and given the increasing sequence database sizes requires long computation times. Here, a single-sequence-based prediction method is presented, called ProteinUnet, leveraging an U-Net convolutional network architecture. It is compared to SPIDER3-Single model, based on long short-term memory-bidirectional recurrent neural networks architecture. Both methods achieve similar results for prediction of secondary structures (both three- and eight-state), half-sphere exposure, and contact number, but ProteinUnet has two times fewer parameters, 17 times shorter inference time, and can be trained 11 times faster. Moreover, ProteinUnet tends to be better for short sequences and residues with a low number of local contacts. Additionally, the method of loss weighting is presented as an effective way of increasing accuracy for rare secondary structures.

KEYWORDS

backbone angles estimation, deep learning, protein structure prediction, secondary structure prediction, solvent accessibility prediction

1 | INTRODUCTION

A three-dimensional protein structure is determined by the amino acid sequences[1,2] and is a key to their functional mechanisms. Experimental determination of the structure is costly and time-consuming compared to sequence determination[3] and the number of known sequences is even 1,000 times bigger than those of examined structures.[4] This creates a need for techniques and models that will computationally predict a protein structure from its primary sequence. The challenge started in 1951 when Pauling and Corey predicted helical and sheet conformations for protein polypeptide backbone[3,5,6] and has not been solved yet.

Accurate protein structure and function prediction rely, in part, on the accuracy of secondary structure prediction that has been extensively studied and resulting in many computational methods (e.g., see an overview[4]). A number of researchers also concentrate on predicting structural properties of proteins like backbone dihedral angles leveraging this information for secondary structure prediction or calculation ([7] based on the early/late-stage protein folding approach[8]).

Recently, developed state-of-the-art methods of secondary structure prediction leverage deep neural network architectures and multiple sequence alignments (MSAs) of homologous sequences allowing them to achieve up to 88% Q3 accuracy,[4] especially for proteins with

a large number of known homologous sequences. However, the majority of proteins do not have any known homologous sequences or very few of them.[9] For such cases, prediction accuracy can deteriorate because of the limited or nonexistent evolutionary information.[10] Moreover, due to the increase in the number of known sequences, the computational time required for finding MSA profiles is also increasing, leading up to multiple hours for longer sequences.

Heffernan et al.[10] took advantage of the recent advancements in deep neural networks and proposed a single-sequence-based model using long short-term memory (LSTM)-bidirectional recurrent neural networks (BRNNs)—SPIDER3-Single. The model can predict multiple one-dimensional (1D) structural properties with relatively high accuracy, especially for nonhomologous sequences.

In this article, we leverage alternative deep neural network architecture—U-Net[11]—for protein structure prediction and compare our results—ProteinUnet—to SPIDER3-Single. The advantage of the U-Net architecture allowed us to reduce the number of parameters in the network and significantly decrease the training and prediction time compared to SPIDER3-Single while maintaining a similar performance of the model. The rest of the article is structured as follows: the second section describes the datasets used in the analysis with a brief description of inputs and outputs used by the models. The next section outlines both algorithms with the description of stratification, weights accounting for rare classes, and training procedures with ensembling. The following section describes the evaluation metrics for classification and regression. Finally, the last two sections present the results and conclusions.

2 | METHODS

2.1 | Datasets

In order to compare our implementation of the SPIDER3-Single model to the original one, we have used the same datasets that were used by the authors of SPIDER3-Single.[10] The original dataset was downloaded from CullPDB[12,13] in February 2017 and split into several smaller datasets. Two of them: TR9993 and TS1199 are listed on the authors' website (https://sparks-lab.org/publication/). Train set TR9993 consists of 9,993 different chains from 9,622 proteins, and test set TS1199 consists of 1,199 chains from 1,187 different proteins. However, 16 of these proteins are no longer available in Protein Data Bank[14] (checked on March 15, 2020). Additionally, 16 chains longer than 1,024 residues were removed from the training set since it is the maximum supported sequence length for ProteinUnet. Thus, we created subsets TR9961 (9,961 chains from 9,592 proteins) and TS1197 (1,197 chains from 1,186 proteins). Also, the performance was tested on 152 proteins from the CASP13 dataset.[15]

2.1.1 | Inputs

The input to the model for a given sequence is a one-hot vector of size $20 \times L$, where L is the length of the protein chain, like in the original article. No other features, like physiochemical properties,[16] BLOSUM matrix,[17] PSSM,[18] nor HHBlits[19] were used. The idea behind the SPIDER3-Single model was to let the neural network learn all the relationships directly from the sequence. The distribution of the amino acids is uneven and ranges from 9.6% for the most common leucine to 1.2% for the rarest cysteine (CYS).

2.1.2 | Outputs

The model outputs could be divided into two main categories: classification and regression outputs. During the classification, the model predicts the secondary structure for eight and three states. The eight states are specified by the secondary structure assignment program Define Secondary Structure of Proteins[20] as follows: there are three helix states: 310-helix (G), alpha-helix (H), and pi-helix (I); three strand states: beta-bridge (B) and beta-strand (E); and three coil types: high curvature loop (S), beta-turn (T), and coil (C). These eight classes are also converted into simpler, three-class problem by grouping the states: G, H, and I into H; B and E into E; and S, T, and C into C. Each problem has separate output nodes in the neural network, resulting in 11 classification output nodes. The distribution of output classes is not even in our datasets. For the eight-class problem, the share of classes ranges from 1% for the rarest I and B classes to 34% for the most common H class. The distributions are very similar between TR9961 and TS1197 datasets.

The regression outputs were calculated using Biopython package[21] and represent accessible surface area (ASA),[22] angles ϕ, ψ, θ, τ (all angles have sine and cosine outputs to remove the effect of the angle's periodicity), half sphere exposure (HSE; there are separate outputs for HSE-up and HSE-down),[23] and contact number (CN). For details, please refer to the original study.[10] In overall, there are 12 regression outputs.

2.2 | Models

All the methods were implemented in the environment containing Python 3.7, TensorFlow 2.2[24] with Keras[25] accelerated by CUDA 10.1, and cuDNN 7.6. (The prediction server based on our ProteinUnet is published on CodeOcean platform (https://codeocean.com/capsule/2521196/tree/v1).)

2.2.1 | SPIDER3-Single

SPIDER3-Single[10] is a network containing two BRNN layers of LSTM units with 256 nodes per direction, followed by the fully connected classifier with two hidden layers with 1,024 and 512 units. LSTM units[26] are used to learn both short and distant dependencies within sequences, and the classifier is used to infer the output from these dependencies.

The input of SPIDER3-Single model is a one-hot encoded single sequence of amino acids. This is the key difference from the original

SPIDER3 model[27] where additional evolutionary features are used like PSSM[18] and HHBlits[19] that are computationally expensive to obtain. Moreover, SPIDER3-Single follows the famous postulate of Anfinsen[1] that the secondary structure of a protein is completely determined by its amino acid sequence alone.

The sizes and activations of the output layers differ between the tasks. For classification, there are two 1-hot encoded output layers of size $3 \times L$ (Q3 output) and $8 \times L$ (Q8 output) followed by softmax activations. For regression, there are four output layers of size $2 \times L$ (sine and cosine for each ϕ, ψ, θ, τ angle) and four output layers of size $1 \times L$ (ASA, CN, HSE α-up, and HSE α-down features) followed by sigmoid activations. The values of the latter output features were normalized to the range <0, 1> by dividing them by their maximum values over the whole training dataset (ASA: 330, CN: 131, HSE α-up: 76, HSE α-down: 79). Additionally, the loss weights for these outputs were set to 2 in order to equalize the contributions of each feature in the loss.

SPIDER3-Single network has nearly 3.2 million trainable parameters from which two-third belong to BRNN part and one-third to the classifier part. This kind of network was proven to be very effective in secondary structure prediction,[27] natural language processing,[28] brain signals analysis,[29] and series forecasting.[30]

In the original study of SPIDER3-Single, the authors presented results of the model repeatedly stacked in the process called iterative learning. Iterative learning significantly increases the training time and complexity giving only small improvements to the accuracy. For purposes of our comparisons with ProteinUnet, we decided not to use iterative learning.

2.2.2 | ProteinUnet

Our 1D fully convolutional ProteinUnet deep neural network consists of a series of blocks placed symmetrically as contractive and expanding paths (that can be broadly thought of as an encoder and decoder), yielding a U-shape.[11] It is a state-of-the-art architecture in the domain of image segmentation.[31,32] The secondary structure prediction for 1D sequences is analogous to the multi-label segmentation of 2D images, but to the best of our knowledge, U-Net architecture has not been used previously for protein structure prediction.[4,33,34] In our proposed architecture, each block in the contractive path contains three convolutional layers with zero padding and kernels of size 7 with stride 1, followed by a rectified linear unit (ReLU) activation. The first two blocks contain 64 filters per layer, and the second two contain 128 filters per layer. Each block ends with an average pooling layer with a kernel of size 2 to perform downsampling.

In the expanding path, there are only two convolution layers per block. Each block is concatenated with the depth-matched block from the contractive path, and then upsampled and passed to the next block. In this manner, high-level features, extracted in the contractive path, propagate through higher-resolution layers of the expanding path. It provides the local context to the global information while upsampling, increasing the precision of the output sequences. Finally, fully connected layers with 128 and 64 ReLU-activated nodes are added as a classifier, followed by an output layer with softmax (for classification network) or sigmoid (for regression network) activations. The architecture diagram of the classification network is presented in Figure 1.

To decrease the number of the parameters and increase the correlation between Q8 and Q3 predictions, the output layer for states Q3 is calculated based on the output for Q8 (unlike in SPIDER3-Single where the outputs for Q8 and Q3 are parallel). All the losses and metrics of the ProteinUnet are the same as in SPIDER3-Single. The total number of trainable parameters of our ProteinUnet classification network is close to 1′597 k which is two times less than for SPIDER3-Single. These two networks have very different hyperparameters (e.g., numbers of filters instead of hidden state dimensions), so they cannot be easily compared. Nevertheless, the training of ProteinUnet is more than 11 times faster, and the inference is over 17 times faster using Tesla K80 GPU, Intel Xeon 2.3 GHz, and 14 GB RAM, as presented in Table 1. Besides having two times fewer parameters, ProteinUnet, being a CNN, benefits more from cuDNN acceleration.[35] Also, a constant size of inputs and outputs in ProteinUnet (in contrast to varying lengths in SPIDER3-Single) makes it easier to implement and manage the memory on GPU.

On the other hand, the constant input size is problematic in terms of variable-length amino acid sequences. Thus, we decided to limit the length of supported sequences in our solution to 1,024 and fill shorter sequences with zeros, masking the loss and metrics accordingly (so the zeros do not affect the results of training or validation).

ProteinUnet, like any other convolutional neural network, processes input sequences as separate patches using a window of the width of the convolutional kernel. Unlike BRNN, it is not sensitive to the order of the timesteps beyond a local scale. However, to recognize more distant patterns, many convolutional layers are stacked with pooling layers, extracting the information from long chunks of the sequence. The receptive field of our ProteinUnet was calculated (using ref. [36]) to be of 710 residues long, so any more distant contacts are impossible to be analyzed. However, such long-range interactions are extremely rare and are present in less than 0.02% residues in our TS1197 dataset.

2.2.3 | Handling imbalanced structure states

Some secondary structure states are relatively rare (like B, G, or I, each present for less than 5% of residues in the Q8 training set) what makes the dataset heavily imbalanced. Interestingly, this issue was not addressed in any previous work.[10,27,33,34] Our solution uses two methods to address this problem: stratification of folds, and adjusting Q8 loss weights to the frequency of the secondary structure states.

There were nine factors of stratification of the training set: the sequence length—shorter/longer than mean sequence length, and one factor for each of eight states occurrence—fewer/more occurrences

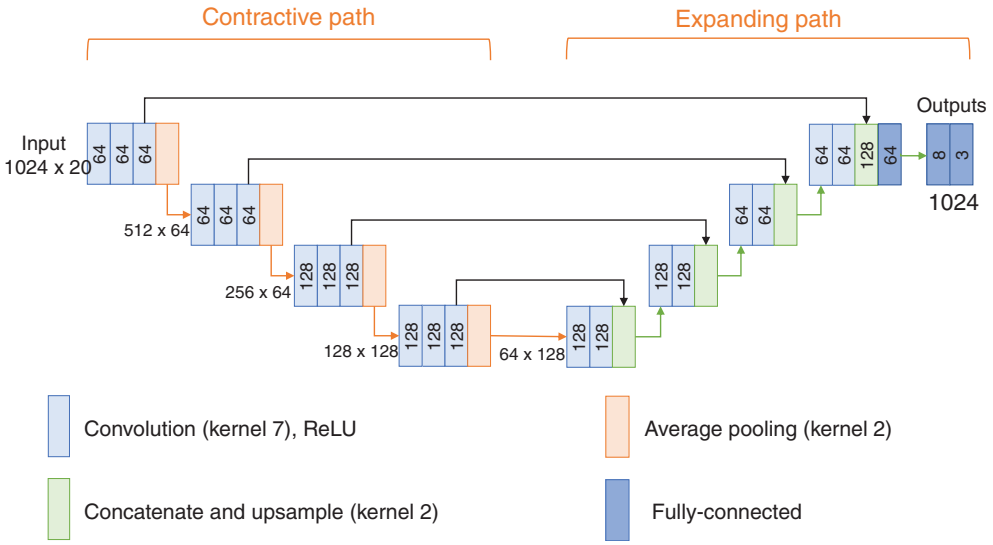

FIGURE 1 The architecture of ProteinUnet secondary structure classification network. The regression network differs only in the number and activations of output layers [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 1 Comparison of mean training and prediction times for SPIDER3-Single and ProteinUnet 10-model ensembles

	Classification		Regression	
	SPIDER3-Single	ProteinUnet	SPIDER3-Single	ProteinUnet
Mean training time per epoch (s)	524.9 ± 1.7	42.0 ± 0.1	527.8 ± 1.7	45.9 ± 0.3
Mean prediction time per chain in TS1197 (s)	1.12 ± 0.54	0.062 ± 0.0025	1.13 ± 0.54	0.066 ± 0.0031

than a mean number of occurrences per chain (C—44.7, H—77.2, E—50.9, T—25.3, G—8.9, S—18.7, I—1.2, B—2.5). This technique ensures that in each of 10 folds there will be a similar ratio of each state. The same stratification was used for both ProteinUnet and SPIDER3-Single.

In a separate section of the article, the method of loss weighting was assessed for ProteinUnet Q8 classification. The weights for four least frequent structures (G, S, I, B) were adjusted to be inversely proportional to the percentage of their occurrence r in the TR9961 dataset using the formula log(0.25/r). This should make ProteinUnet to pay more attention to the rare states.

2.2.4 | Training procedures and ensembling

The training dataset was divided into 10 stratified folds for cross-validation. For a fair comparison, both architectures were trained using the same division into folds. Each of 10 models was trained using Adam optimizer[37] with batch size 8 and initial learning rate 0.001. Early stopping condition was used when the validation loss was not improving for 5 epochs. The training lasted from 12 to 16 epochs for classification (ProteinUnet—M = 13.4, *SD* = 0.9; SPIDER3-Single—M = 13.9, *SD* = 1.1) and from 13 to 20 epochs for regression (ProteinUnet—M = 15.7, *SD* = 2.2; SPIDER3-Single—M = 15.5, *SD* = 1.1). After the training, the ensemble was created from all the 10 models by taking the average of their outputs, forming the final prediction on the test set.

There is no information about a batch size or a learning rate in articles about SPIDER3[27] nor SPIDER3-Single.[10] Due to the variable length of the input of SPIDER3-Single, the training with a batch size of 8 was implemented in a way where all the sequences in the batch are filled with zeros up to the length of the longest sequence in the batch. The loss and metrics are masked accordingly, so these additional zeros do not affect the results of training or validation. All the predictions on the test set were performed with a batch size 1 (one-by-one, without zero padding).

2.3 | Evaluation metrics

2.3.1 | Classification

The simplest and most popular measures of protein secondary structure prediction quality are average three-state per-residue accuracy Q3 and eight-state per-residue accuracy Q8. They give the percentage of residues for which the predicted secondary structures are correct[38,39] according to Equation (1)

$$Q_m = 100\% \times \frac{\sum_{i=1}^{m} M_{ii}}{N_{res}} \quad (1)$$

where m is the number of classes, N_{res} is the total number of residues, and M_{ii} is the number of correctly predicted residues in state i. Q3 and

Q8 accuracies are defined for $m = 3$ and $m = 8$, respectively.[40] Since Q3 and Q8 are reported in almost every article, including the original SPIDER3-Single study, we will use them in our comparisons as well.

2.3.2 | Regression

The continuous variables are split into two groups, following the methodology described by SPIDER3-Single authors,[10] and each of them measures performance differently. ASA, CN, HSEα-up, and HSEα-down predicted values are compared to the true values using the Pearson correlation coefficient (CC), defined as Equation (2)

$$CC = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad (2)$$

where n is the sample size, x_i and y_i are the individual sample points indexed with i, x is the sample mean for the x variable, and y is the sample mean for the y variable.

The performance of the ϕ, ψ, θ, and τ angles are calculated as the circular mean absolute error, which is the smaller of α_i and $(360^\circ - \alpha_i)$ to account for the periodicity of the angles, where $\alpha_i = \left|\alpha_i^{pred} - \alpha_i^{true}\right|$, α_i^{pred} is the predicted angle value, and α_i^{true} is the true angle value.

3 | RESULTS

The comparison of overall results on the test sets between the original SPIDER3-Single Iteration 2 (authors do not report results for Iteration 1), our reimplementation of SPIDER3-Single, and the new proposed ProteinUnet is presented in Table 2. Because of all mentioned differences, it is impossible to directly compare the original and reimplemented SPIDER3-Single. However, the results are on the similar level. In the direct comparison to the reimplemented SPIDER3-Single, our ProteinUnet achieved better classification accuracies, but worse results for angles. However, all the differences are smaller than 2%.

Table 3 shows the mean accuracies of Q3 and Q8 predictions at the sequence level in TS1197 and CASP13, along with *SD*s, and *p*-values of the two-sided Wilcoxon signed-rank test between models. For TS1197, ProteinUnet gives better mean accuracies and lower *SD*s than SPIDER3-Single. The difference for Q3 is significant at $p < .05$, and for Q8 at $p < .0001$. For CASP13 dataset, ProteinUnet gives worse results for Q3 ($p < .05$), and very similar results for Q8 ($p = .90$).

3.1 | Classification

3.1.1 | Analysis per amino acid

The analysis of the classification accuracy per amino acid type is presented in Figure 2. The rare amino acids tend to have worse accuracy, like CYS, histidine (HIS), or tryptophan (TRP). From the rare amino acids, only methionine has accuracy above the average. The best Q3 accuracy for both models was achieved for proline (PRO): 76.26% for ProteinUnet and 76.69% for SPIDER3-Single. The biggest difference for Q3 in favor of ProteinUnet is for TRP—0.48 pp. and in favor of SPIDER3-Single for tyrosine—0.46 pp.

Surprisingly, the Q8 accuracy for PRO is below average, and the best performing Q8 amino acid is isoleucine (ILE). Similarly, the worst Q8 accuracy was achieved for glycine (GLY) which shows above average results for Q3. The biggest difference for Q3 in favor of ProteinUnet is for CYS—0.90 pp. and in favor of SPIDER3-Single for PRO—0.69 pp.

TABLE 2 The comparison of performance for test sets between (a) original SPIDER3-Single Iteration 2,[10] (b) our reimplementation of SPIDER3-Single, and (c) ProteinUnet according to fraction of residues in correctly predicted three and eight states (Q3 and Q8), Pearson CC, and MAE

	(a)	(b)	(c)
	TS1199	TS1197	TS1197
Q3	72.56%	72.56%	72.66%
Q8	60.11%	59.88%	60.06%
ASA (CC)	0.671	0.669	0.667
HSEα-up (CC)	0.612	0.608	0.602
HSEα-down (CC)	0.568	0.566	0.567
CN (CC)	0.643	0.618	0.621
ϕ (MAE)	24.5	23.5	23.7
ψ (MAE)	43.5	41.8	42.3
θ (MAE)	11.3	10.1	10.2
τ (MAE)	45.8	43.2	43.8

Abbreviations: ASA, accessible surface area; CC, correlation coefficients; HSE, half sphere exposure; MAE, mean absolute error.

3.1.2 | Analysis per sequence length

Figure 3 presents the Q3 accuracy as a function of sequence length. The linear regression models show that ProteinUnet has a higher accuracy for shorter chains but its accuracy decreases faster than for SPIDER3-Single with increasing sequence length. The Q3 accuracy of the ProteinUnet was below 40% only for one chain, while for SPIDER3-Single—six chains. Moreover, ProteinUnet achieved 100% Q3 accuracy for one protein sequence (2O6N Chain A) while SPIDER3-Single was never 100% correct.

The biggest difference at the sequence level in favor of ProteinUnet was for protein 1T1V Chain A with 93 residues for which ProteinUnet achieved 79.57% while SPIDER3-Single only 54.84%. SPIDER3-Single had the biggest advantage over ProteinUnet for protein 1KAF Chain A with 108 residues for which ProteinUnet achieved 65.74% while SPIDER3-Single 83.33%. In overall, ProteinUnet achieved better results for 578 sequences while SPIDER3-Single for

TABLE 3 Performance in secondary structure prediction by ProteinUnet and SPIDER3-Single on TS1197 and CASP13[15] according to mean accuracy and *SD* at the sequence level

		TS1197			CASP13		
		Mean (%)	*SD* (%)	*p*-Value	Mean (%)	*SD* (%)	*p*-Value
Q3	ProteinUnet	73.53	8.70	.0152	74.39	8.13	.0128
	SPIDER3-Single	73.18	9.04		75.12	7.65	
Q8	ProteinUnet	61.82	10.86	<.0001	60.81	12.17	.8961
	SPIDER3-Single	61.34	11.15		60.81	12.79	

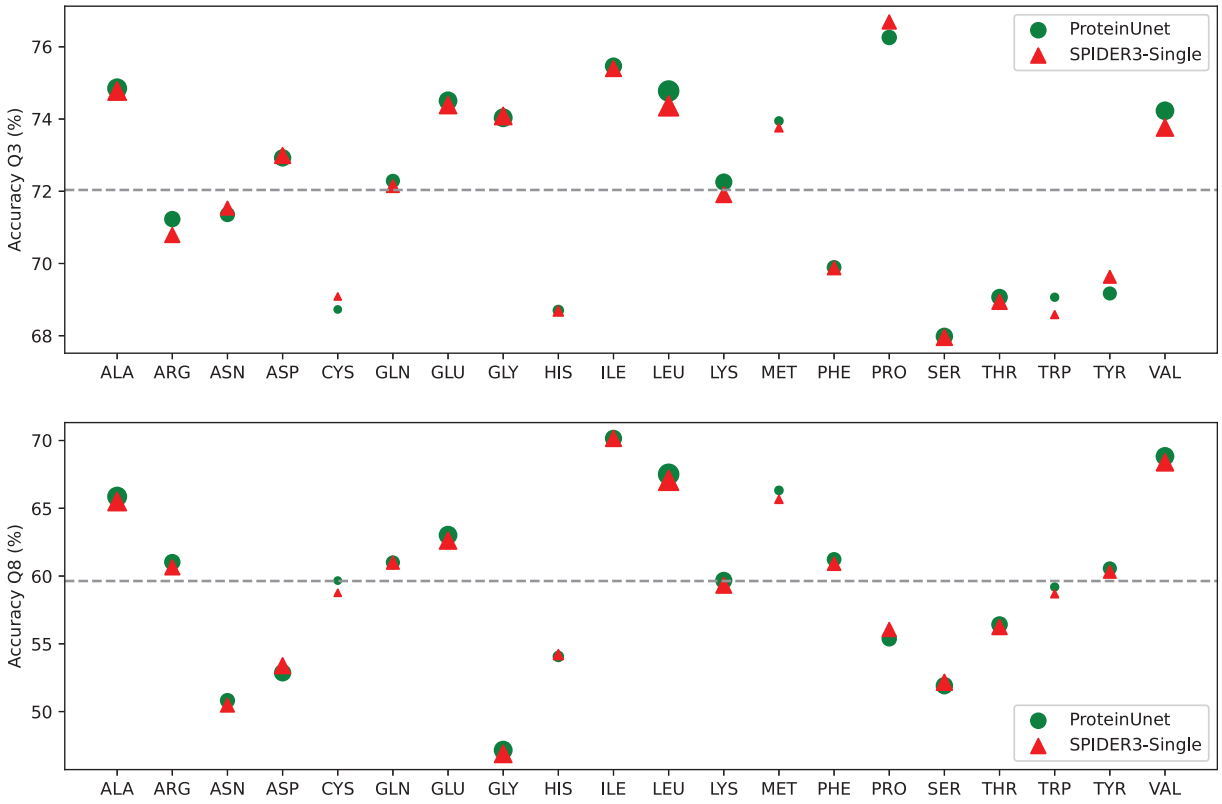

FIGURE 2 Accuracy of the secondary structure prediction (Q3 and Q8) for individual amino acids for SPIDER3-Single (red triangles) and ProteinUnet (green circles) on TS1197 dataset. Three-letter codes were used for amino acid residues. The size of the bubble represents the frequency of the amino acids. The gray horizontal line marks the fraction of residues in correctly predicted three and eight states (Q3 and Q8) [Color figure can be viewed at wileyonlinelibrary.com]

520, respectively. For 99 sequences, both models achieved the same results, which do not necessarily mean they had the same predictions since the mistakes might have been on different positions.

3.1.3 | Influence of Q8 loss weighting

The results for ProteinUnet with weighted Q8 loss are presented in Figure 4 in comparison to nonweighted versions of ProteinUnet and SPIDER3-Single. As expected, weighting helped to achieve much better accuracies for all rare states (G, S, I, B). The highest increase (by 9 pp.) was noticed for structure G (3_{10}-helix). For Structures B (beta-bridge) and I (pi-helix), weighting allowed to pull the accuracy out of 0% level. As a side effect of weighting, for states C (coil), H (alpha-helix), and T (beta-turn) accuracies decreased up to 2 pp. and were lower than for nonweighted ProteinUnet and SPIDER3-Single. This caused the overall Q8 accuracy for weighted ProteinUnet to be slightly worse than before weighting, at both sequence (61.59%) and residues level (59.83%). Interestingly, after weighting, the accuracy for a frequent E state (beta-strand) was better than for nonweighted ProteinUnet and SPIDER3-Single. All the effects mentioned in this section were statistically significant at $p < .005$ according to the two-sided Wilcoxon signed-rank tests.

3.2 | Regression

Figure 5 presents the distribution of the regression outputs for TS1197 dataset. The majority of ϕ angles are close to the -63° and the predictions of both models are most common for -65°. However,

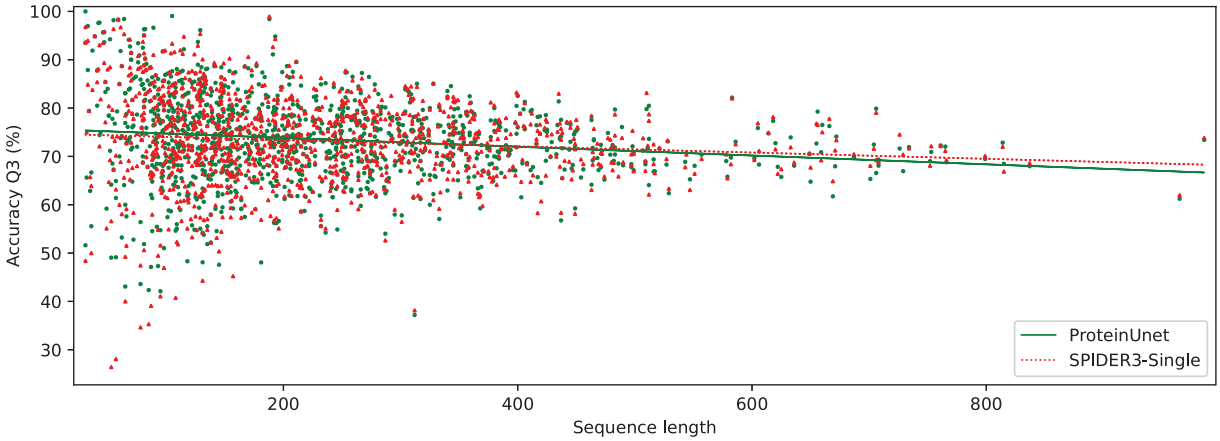

FIGURE 3 The accuracy of secondary structure prediction (Q3) for individual sequences against the sequence length for ProteinUnet (green circles) and SPIDER3-Single (red triangles) on TS1197 dataset [Color figure can be viewed at wileyonlinelibrary.com]

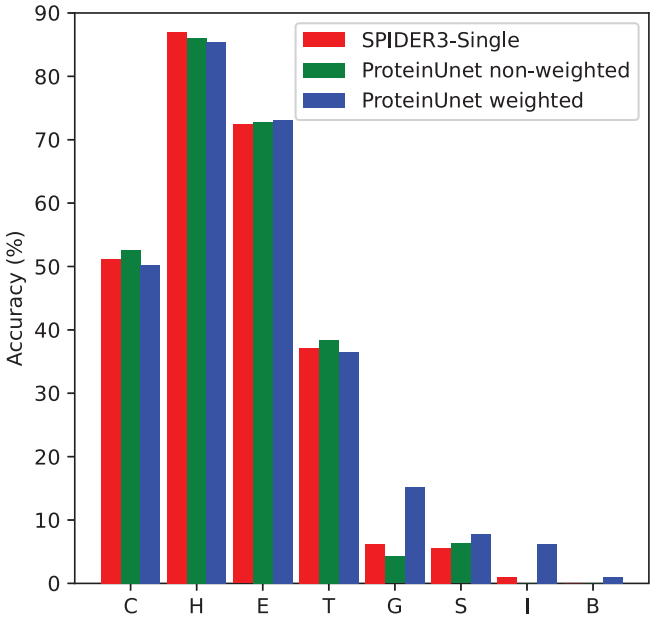

FIGURE 4 The comparison of mean accuracy at the sequence level for each Q8 state on TS1197 dataset between weighted and nonweighted ProteinUnet and nonweighted networks [Color figure can be viewed at wileyonlinelibrary.com]

both models rarely predict values below −125°, but more than 40,000 residues have true values below −125°. The ψ angles are grouped around two local maxima: −42 and 135°. Surprisingly, the predictions for ψ below −45° or above 150° are rare, while in true values, they account for more than 73,000 cases.

The majority of θ and τ angles are close to the 91° maximum and around 117° local maximum and the values span from 64 to 177°. Both models' predictions fall between 84 and 145°, so the long tails are not predicted at all. Moreover, values between 95 and 120° were predicted much more often than they occurred. The τ angle predictions are grouped around two local maxima: 50 and −165°, but the true values are more distributed. Especially, the predictions around −165° are more than two times often than they actually occur. For angle τ prediction, SPIDER3-Single tends to predict more often the values around maxima than ProteinUnet. Both models fail to predict the cases when ASA is equal to 0 with SPIDER3-Single predictions slightly shifted to lower values. The sigmoid output function might be the reason for the poor performance of the predictions around 0 value. The ASA predictions for both models do not exceed 190, while the maximum true value was 297. The CN values span from 0 to 84, while the model predictions range between 10 and 65. SPIDER3-Single prediction distribution is shifted to higher values. The distribution of HSE α-up predictions does not resemble the true value distribution. The maximum predicted value was 31, while the maximum true value was 45. The true values of HSE α-down range between 2 and 51, while the predictions fall between 6 and 35. For both HSE, predictions from SPIDER3-Single are shifted more toward higher values compared to ProteinUnet.

3.3 | Local contacts analysis

Figures 6 and 7 show the dependence of the accuracy of secondary structure prediction on the number of local and nonlocal contacts in a residue, respectively. Exactly like in ref. [10], nonlocal contacts are defined as contacts between two different residues that are more than or equal to 20 residues away in their sequence positions, but less than 8 A° away in terms of their atomic distances between Cα atoms. Each point presented on the plots has a representation of at least 1,000 residues. For both ProteinUnet and SPIDER3-Single, accuracy for Q3 decreases sharply with the number of local and nonlocal contacts greater than 2. ProteinUnet shows noticeably better results for residues with a small number of local contacts (<3), but noticeably worse results for those with more than nine nonlocal contacts. It confirms that ProteinUnet is better at capturing close local dependencies (up to 12 pp. more for two local contacts), but worse at analyzing long-range interactions (up to 2 pp. less for 11 nonlocal contacts). The number of nonlocal contacts is correlated with the length of the sequence, so it may partly explain the trend visible in Figure 3.

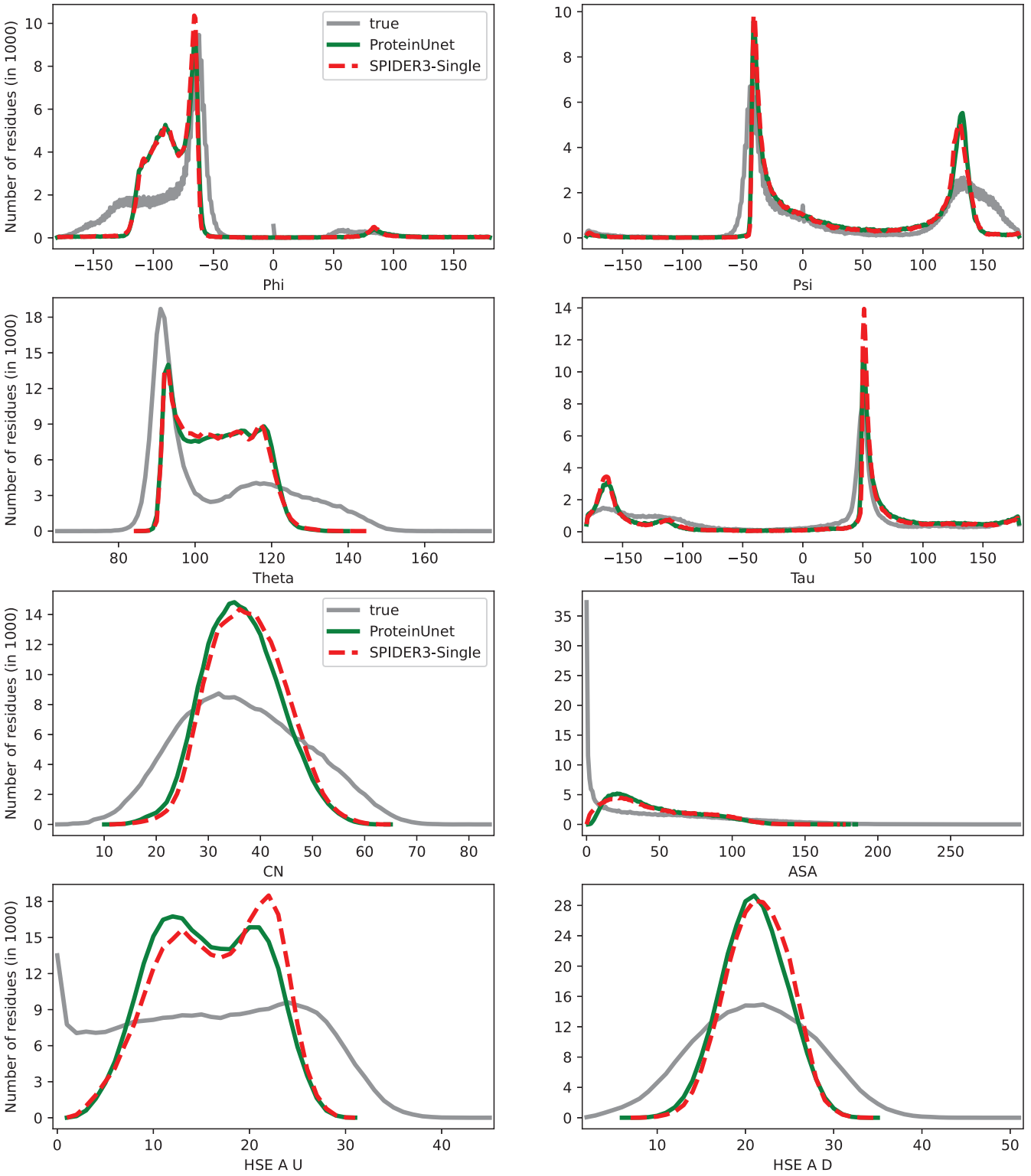

FIGURE 5 The distribution of regression outputs for TS1197 dataset. True values are presented with a solid gray line, prediction values for ProteinUnet with a solid green line and SPIDER3-Single with a dashed red line [Color figure can be viewed at wileyonlinelibrary.com]

4 | CONCLUSIONS

ProteinUnet is the first model that successfully leverages U-Net deep learning architecture for sequence-based protein 1D structural properties prediction. The model does not use the evolutionary profiles generated from MSA like PSSM or HHblits, which are computationally expensive to calculate. It achieves comparable results to state-of-the-art sequence-based model—

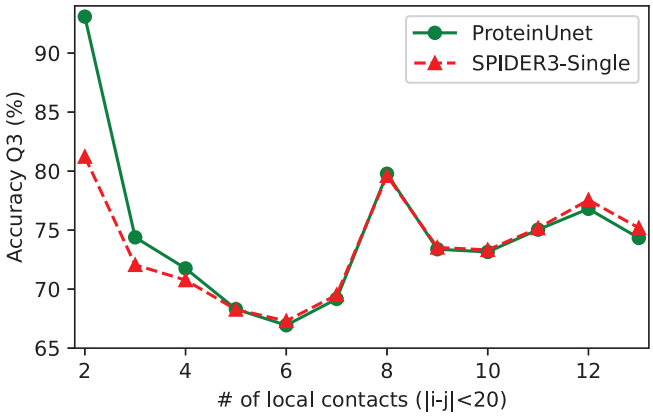

FIGURE 6 Accuracy of predicted Q3 as a function of the number of local contacts on TS1197 dataset [Color figure can be viewed at wileyonlinelibrary.com]

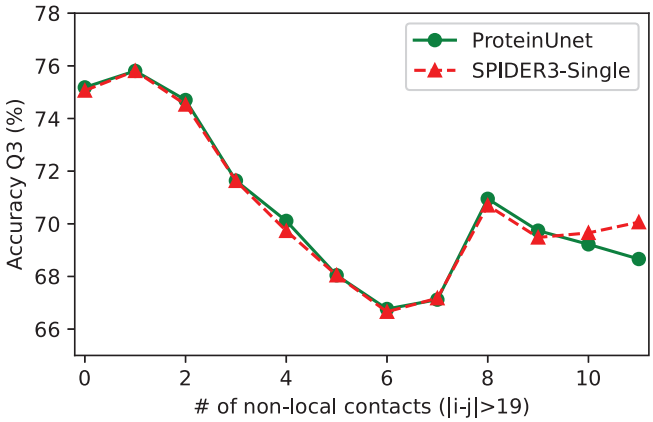

FIGURE 7 Accuracy of predicted Q3 as a function of the number of nonlocal contacts on TS1197 dataset [Color figure can be viewed at wileyonlinelibrary.com]

SPIDER3-Single based on LSTM-BRNN architecture while having two times fewer parameters and running several times faster (11 times faster training and 17 times faster inference). It makes it especially useful in large-scale predictions and applications on low-cost and embedded devices. Moreover, ProteinUnet shows better results for short sequences and residues with a low number of local contacts, so should be used preferably to SPIDER3-Single when these factors matter. Additionally, our experiments showed that the proposed weighting procedure can be effectively used in ProteinUnet to substantially increase the accuracy on the rare states. The results on CASP13 dataset confirm that ProteinUnet performs as good as SPIDER3-Single for completely untrained folds.

The disadvantages of the proposed architecture are mainly connected with the limited receptive field of convolutional networks. They include decreased accuracy for long chains and residues with many nonlocal contacts. However, they may be addressed in the future by increasing the context or receptive field of U-Net, or adding iterative training as described in ref. [10]. Moreover, the next future step is to improve the weighting procedure to avoid the decrease on the more frequent states.

ORCID

Katarzyna Stapor https://orcid.org/0000-0003-3003-6592

REFERENCES

[1] C. B. Anfinsen, *Science* **1973**, *181*, 223. https://doi.org/10.1126/science.181.4096.223.
[2] B. Rost, C. Sander, R. Schneider, *J. Mol. Biol.* **1994**, *235*, 13. https://doi.org/10.1016/S0022-2836(05)80007-5.
[3] Y. Yang, J. Gao, J. Wang, R. Heffernan, J. Hanson, K. Paliwal, Y. Zhou, *Brief. Bioinform.* **2018**, *19*, 482. https://doi.org/10.1093/bib/bbw129.
[4] T. Smolarczyk, I. Roterman-Konieczna, K. Stapor, *Curr. Bioinf.* **2020**, *15*, 90.
[5] L. Pauling, R. B. Corey, H. R. Branson, *Proc. Natl. Acad. Sci. U. S. A.* **1951**, *37*, 205. https://doi.org/10.1073/pnas.37.4.205.
[6] L. Pauling, R. B. Corey, *Proc. Natl. Acad. Sci. U. S. A.* **1951**, *37*, 729. https://doi.org/10.1073/pnas.37.11.729.
[7] T. Smolarczyk, K. Stapor, I. Roterman-Konieczna, *Bio-Algorithms Med-Syst.* **2019**, *15*. https://www.degruyter.com/view/journals/bams/15/4/article-20190034.xml.
[8] I. Roterman-Konieczna, *Protein Folding in Silico*, 1st ed., Woodhead Publishing, Cambridge **2012**.
[9] S. Ovchinnikov, H. Park, N. Varghese, P.-S. Huang, G. A. Pavlopoulos, D. E. Kim, H. Kamisetty, N. C. Kyrpides, D. Baker, *Science* **2017**, *355*, 294.
[10] R. Heffernan, K. Paliwal, J. Lyons, J. Singh, Y. Yang, Y. Zhou, *J. Comput. Chem.* **2018**, *39*, 2210.
[11] O. Ronneberger, P. Fischer, T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," arXiv, vol. 1505.04597, no. cs. CV, **2015**.
[12] G. Wang, R. L. Dunbrack Jr.., *Bioinformatics* **2003**, *19*, 1589.
[13] G. Wang, R. L. Dunbrack Jr.., *Nucleic Acids Res.* **2005**, *33*, W94.
[14] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, *Nucleic Acids Res.* **2000**, *28*, 235. https://doi.org/10.1093/nar/28.1.235.
[15] A. Senior, R. Evans, J. John, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, Z. Augustin, A. Nelson, *Proteins* **2019**, *87*, 1141.
[16] J. Fauchère, M. Charton, L. B. Kier, A. Verloop, V. Pliska, *Int. J. Pept. Protein Res.* **1988**, *32*, 269.
[17] S. Henikoff, J. G. Henikoff, *Proc. Natl. Acad. Sci. U. S. A.* **1992**, *89*, 10915.
[18] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman, *Nucleic Acids Res.* **1997**, *25*, 3389.
[19] M. Remmert, A. Biegert, A. Hauser, J. Söding, *Nat. Methods* **2012**, *9*, 173.
[20] W. Kabsch, C. Sander, *Biopolymers* **1983**, *22*, 2577.
[21] P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, M. J. L. de Hoon, *Bioinformatics* **2009**, *25*, 1422.
[22] C. Chothia, *Nature* **1974**, *248*, 338.
[23] T. Hamelryck, *Proteins: Struct. Funct. Bioinf.* **2005**, *59*, 38.
[24] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, R. Jozefowicz, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems," [Online]. http://tensorflow.org/
[25] F. Chollet "Keras [Online]," **2015**. https://keras.io/getting_started/faq/#how-should-i-cite-keras.
[26] S. Hochreiter, J. Schmidhuber, *Neural Comput.* **1997**, *9*, 1735.
[27] R. Heffernan, Y. Yang, K. Paliwal, Y. Zhou, *Bioinformatics* **2017**, *33*, 2842.

[28] Y. Cai, Q. Huang, Z. Lin, J. Xu, Z. Chen, Q. Li, *Knowl. Based Syst.* **2020**, *203*, 105856.
[29] A. Craik, Y. He, J. L. Contreras-Vidal, *J. Neural Eng.* **2019**, *16*, 031001.
[30] K. Bandara, C. Bergmeir, S. Smyl, *Expert Syst. Appl.* **2020**, *140*, 112896.
[31] J. Nalepa, P. R. Lorenzo, M. Marcinkiewicz, B. Bobek-Billewicz, P. Wawrzyniak, M. Walczak, M. Kawulok, W. Dudzik, K. Kotowski, I. Burda, B. Machura, G. Mrukwa, P. Ulrych, Michael, *Artif. Intell. Med.* **2020**, *102*, 101769.
[32] K. Kotowski, J. Nalepa, W. Dudzik, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, Springer International Publishing, Cham **2020**.
[33] W. Wardah, M. Khan, A. Sharma, M. A. Rashid, *Comput. Biol. Chem.* **2019**, *81*, 1.
[34] J. Hanson, K. K. Paliwal, T. Litfin, Y. Yang, Y. Zhou, *J. Comput. Biol.* **2020**, *27*, 796.
[35] T. Lei, Y. Zhang, S. I. Wang, H. Dai, Y. Artzi, "Simple Recurrent Units for Highly Parallelizable Recurrence," arXiv, vol. cs.CL, no. 1709.02755, **2017**.
[36] A. Araujo, W. Norris, J. Sim, *Distill* **2019**, *4*. https://distill.pub/2019/computing-receptive-fields/.
[37] D. P. Kingma, J. Ba, "Adam: A Method for Stochastic Optimization," arXiv, vol. cs.LG, no. 1412.6980, **2014**.
[38] Q. Jiang, X. Jin, S.-J. Lee, S. Yao, *J. Mol. Graphics Modell.* **2017**, *76*, 379. https://doi.org/10.1016/j.jmgm.2017.07.015.
[39] S. Wang, J. Peng, J. Ma, J. Xu, *Sci. Rep.* **2016**, *6*. https://www.nature.com/articles/srep18962.
[40] J. Lee, *Proteins* **2006**, *65*, 453. https://doi.org/10.1002/prot.21164.

How to cite this article: Kotowski K, Smolarczyk T, Roterman-Konieczna I, Stapor K. ProteinUnet—An efficient alternative to SPIDER3-single for sequence-based prediction of protein secondary structures. *J Comput Chem*. 2021;42:50–59. https://doi.org/10.1002/jcc.26432